Measurably evolving pathogens in the genomic era

R Biek[1,2], O G Pybus[3], J O Lloyd-Smith[2,4], X Didelot[5]

[1] Institute of Biodiversity, Animal Health and Comparative Medicine, Boyd Orr Centre for Population and Ecosystem Health, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

[2] Fogarty International Center, National Institutes of Health, Bethesda MD, USA

[3] Department of Zoology, University of Oxford, Oxford, UK

[4] Department of Ecology and Evolutionary Biology, University of California at Los Angeles, Los Angeles, CA 90095, USA

[5] Department of Infectious Disease Epidemiology, Imperial College London, UK

Highlights

- Pathogens are measurably evolving if substantial evolutionary change is detectable between genetic samples taken at different times
- Whole-genome sequencing has massively increased the range of measurably evolving pathogens
- In the future we expect novel and important insights into pathogen population dynamics to come from genomic data
- Considerable analytical challenges will need to be overcome to fully realise this potential

Keywords: bacteria; DNA virus, epidemiological models; evolutionary rate; infectious disease; phylodynamics

Abstract

Current sequencing technologies have created unprecedented opportunities for studying microbial populations. For pathogens with comparatively low per-site mutation rates, such as DNA viruses and bacteria, whole-genome sequencing can reveal the accumulation of novel genetic variation between population samples taken at different times. The concept of “measurably evolving populations” and related analytical approaches have provided powerful insights for fast-evolving RNA viruses, but their application to other pathogens is still in its infancy. Here we argue that previous distinctions between slow- and fast-evolving pathogens become blurred once evolution is assessed at a genome-wide scale and we highlight important analytical challenges to be overcome in order to infer pathogen population dynamics from genomic data.

A changing landscape for studying pathogen evolution

Over a decade ago, Drummond *et al.* [1] introduced the idea of "measurably evolving populations". These populations exhibit detectable amounts of *de novo* evolutionary change among genetic sequences sampled at different time points. This concept, and the analytical methodology it has spawned, have revolutionised our ability to study population dynamic processes using genetic sequence data (Box 1). Until recently, RNA viruses were the primary target of such approaches owing to their high per-site evolutionary rates, one consequence of which being that partial genome sequences accumulate observable genetic change on the same time scales at which epidemiological processes occur. Using analytical techniques derived from population genetic theory, aspects of these processes can be inferred from sequence data. The application of such techniques to RNA viruses has been a highly successful scientific endeavour, giving rise to the field of phylodynamics [2–4] and yielding many fundamental insights about infectious disease epidemiology [5].

In contrast to RNA viruses, pathogens that evolve more slowly per nucleotide site, such as bacteria and double stranded (ds) DNA viruses, have not been considered amenable to these approaches – until recently. Because mutation rates and genome sizes tend to be inversely related (Figure 1), slow-evolving pathogens can accumulate novel variation throughout their generally larger genomes on a time scale similar to that seen in RNA viruses, and sometimes on time scales similar to relevant epidemiological processes. With the rise of novel sequencing technologies it has become increasingly feasible to routinely sequence whole genomes of a diverse range of microbes. This has pushed many new pathogen systems into the realm of measurably evolving populations, offering opportunities to gain insights into their epidemiology and population dynamics [6–8].

While we share the excitement about these prospects we argue that substantial challenges must be overcome once the tools and concepts that have, until now, been tested on RNA viruses become applied to other pathogens. Here, we provide an overview of issues and research problems that arise in this emerging field by focussing on three main areas: (i) the relative time scale of evolutionary and epidemiological processes; (ii) the effect of temporal sampling scale on evolutionary rate estimates; (iii) the novel analytical challenges arising from the biological characteristics of bacterial and DNA viral pathogens. We focus on these two groups because they are an increasingly common target for whole genome studies, but applications to other pathogens, such as fungi, are also starting to become available [8]. Throughout this article, we discuss the need for novel approaches and suggest possible ways forward.

Measuring evolution on epidemiological time scales

Genomic data from populations sampled through time are being generated for an increasing range of pathogens. As a consequence, estimates of evolutionary rates (nucleotide or amino acid substitutions per site or codon per year) are becoming available for many of these pathogens for the first time. A comparison of genome-wide evolutionary rates (as opposed to *per-site* rates), estimated from pathogens sampled over at least one decade, confirms that the dichotomy between 'fast-evolving' and 'slow-evolving' pathogens is more appropriately viewed as a continuum that does not conform to taxonomic boundaries: many whole bacterial genomes sequenced to date are accumulating novel mutations over time frames of days to months, similar to those typically observed in RNA viruses (Figure 2). For dsDNA viruses, information is more limited but so far suggests that this time frame might be more in the order of months to years [9–11], likely due to a combination of lower per

site mutation rates and smaller genomes. Slow rates of evolution are also seen in certain bacteria such as *Mycobacteria* [12–14].

How quickly genetic variation is expected to arise within the pathogen genome has implications for the scale at which epidemiological processes can be realistically resolved. For example, at the finest epidemiological resolution of direct transmission from host to host, even genome-wide variation might be insufficient to result in distinguishable consensus genomes between donor and recipient hosts, especially where the average generation interval between donors and recipients is short (Figure 2). In some cases, within-host variation that is not represented in the consensus genome, in the form of rare and transient variants, can provide additional information for identifying transmission links [15]. However, the degree to which this applies to bacteria and dsDNA viruses is not clear and few empirical studies so far have examined within-host levels of variation for these pathogen groups. A recent simulation study found that within-host variation was not sufficient to enable accurate mapping of transmission links for bacterial pathogens – and indeed showed that such variation can obscure transmission links if inference is based on single isolates from each host [16]. The reconstruction of transmission pathways from genetic data is a complex problem under the best of circumstances [5,6], hence researchers need to carefully consider for each pathogen whether the epidemiological process of interest occurs at a time scale at which sufficient genetic signal is detectable.

Several research avenues can be pursued to overcome the uncertainty resulting from limited genetic resolution. The simplest is to shift attention to higher hierarchical scales (e.g. towns instead of individuals) resulting in longer average waiting times between transmission events, thereby increasing the probability of successful genetic tracing. Alternatively, genomic variation other than single nucleotide polymorphisms (SNPs) could be used to infer transmission links, such as insertions and deletions (indels) or auxiliary parts of bacterial genomes (e.g. plasmids). For example, a recent

outbreak of the dsDNA monkeypox virus exhibited variation in indels but not in nucleotide sequence [17] and indels have been used to aid the reconstruction of within-host HIV-1 evolution [18]. However, whether this strategy has practical use for revealing the phylogenetic structure of pathogens with limited nucleotide polymorphism remains to be tested. Available bacterial data (e.g. for *Mycobacterium tuberculosis* and *Staphylococcus aureus*) suggest that SNPs outnumber indels, so it is unclear whether indel variation would be common enough to be useful. A third strategy is to integrate genetic data with other types of information such as host contact or incidence patterns, thus maximising the total information available for analysis. This represents a vibrant and rapidly expanding area of research (see Box 2) which might bring substantial future advances in epidemiological inference [15,19–21].

Time-dependency of evolutionary rates and its consequences

The power to detect measurable evolution grows as sequences are sampled further apart in time [1]. However, the length of the time interval over which sequences are sampled can also affect the apparent rate of genetic change estimated from the data, with longer time scales yielding lower evolutionary rate estimates (Figure 3). The causes of this decline in the apparent rate of molecular evolution are not fully understood, but might include changes in selective constraint, nucleotide saturation at variable sites, and the effect of slightly deleterious mutations that are slower to be removed by purifying selection. This ‘time-dependency’ effect was first recognised in data arising from ancient DNA studies of animal species [22] and in recent years evidence has accrued that the same phenomenon also applies to bacteria and viruses [23–25].

Taxonomic information about the timing of host divergence events can be used to calibrate the age of ancestral nodes in pathogen phylogenies (often with considerable uncertainty) and thus to quantify the amount of genetic change over different time frames. A recent analysis of sequence data for primate lentiviruses, representing time scales ranging from months to millennia, revealed that evolutionary rate estimates do indeed decline as longer time intervals are considered [25]. Due to the increasing availability of complete microbial genomes, similar comparisons are now also possible for bacterial pathogens, and confirm the same pattern of apparent rate decline (Figure 3). In general, bacteria tend to be more resistant than most viruses to degradation outside the host and in some instances this has enabled bacterial genomes to be recovered from the bones of infected hosts who died more than a thousand years ago [26–28], providing novel opportunities for the estimation of long-term evolutionary rates without reliance on potentially erroneous divergence time calibrations.

The time-dependency of evolutionary rate estimates has important implications for genetic inference of pathogen population dynamics. On the one hand, estimates based on co-divergence dates will be misleadingly low when considering the amount of divergence expected over epidemiological time scales. Conversely, rates estimated from sequences sampled over short time scales should not be used to determine phylogenetic divergence times in the more distant past, because they will systematically underestimate these ages [24]. For example, early studies of RNA viruses that extrapolated molecular clock results in this way commonly produced unrealistically young ages for their most recent common ancestors [29,30]. Recent approaches have at least partially overcome this problem by using evolutionary models that account for the effect of purifying selection on long-term rates [31,32]. While saturation at synonymous sites still remains a problem in some cases, especially for RNA viruses, selection-aware models have been applied with particular

success to DNA viruses, where saturation can be less pronounced [33]. The same should be true within individual bacterial species, where these methods therefore hold promise but so far remain untested.

Biological complexities

Extending the concept of a measurably evolving population from RNA viruses to DNA viruses and bacteria requires taking into account several biological processes and patterns that are absent or rare in RNA viruses, and for which current analytical tools are insufficient. We highlight three key examples of such complexities that will require theoretical and methodological advances over the coming years.

Intra-genomic and temporal variation in evolutionary rate

When averaged across the genome, the per-site evolutionary rate among different RNA viruses, dsDNA viruses or bacterial species can vary by several orders of magnitude [7,34] (see Figure 1). More importantly from an analytical point of view, there is often considerable rate heterogeneity within species. The evolutionary rate of bacteria, for example, varies significantly along their genomes [35,36]. This variation is partly explained by the fact that mutation is highly regulated by DNA repair mechanisms in a manner that is still not fully understood [37]. There is evidence for a lower rate of mutation in highly expressed genes, which suggests that bacteria have mechanisms to differentially control the frequency of mutations within their genomes [38]. Molecular evolution can also vary temporally, as was recently reported for *Yersinia pestis* and hypothesized to reflect strong demographic fluctuations [39]. On a shorter time scale, modulation of DNA repair pathways can increase the substitution rate dramatically in a process called hypermutation, which has been

shown to be adaptively important, for example to gain antibiotic resistance [40,41]. The life cycles of many infectious bacterial pathogens introduce further possible causes of temporal variation in evolutionary rates. For example, *Mycobacterium tuberculosis* infections can be active or latent, and different rates of evolution have been reported in these two phases [42]. Similarly, a difference between evolutionary rates would be expected for the latent and active replication phases of some DNA viruses, such as herpesviruses [34]. Some bacterial pathogens from the firmicute phylum, such as *Bacillus* or *Clostridium,* produce endospores that can survive for years outside a host, and since these structures do not reproduce they are likely to accumulate fewer mutations on average compared to actively dividing cells [43,44]. Whereas heterogeneity across the genome can be accounted for by partitioning the data for estimating rates, temporal heterogeneity is more challenging to represent using current molecular clock models and can even obscure any signal of measurable evolution (Box 3).

Homologous recombination

Homologous recombination is frequent in the evolution of most bacterial species, and this process is often found to play a greater part in genomic diversification than *de novo* mutation [45]. The recombination rate varies significantly among species, and can vary from one lineage to another within a species [43,46,47] and among genome regions [46,48]. Recombination is non-random in terms of the pairs of donors and recipients involved, with more frequent exchanges happening between close relatives [49] or occupants of the same ecological niche [50,51]. Recombination has also long been known to occur in some large DNA viruses, such as poxviruses, and whole-genome sequencing is providing new opportunities to shed light on its frequency and genomic consequences [52]. Ignoring recombination when analysing

measurably evolving pathogens can be misleading [53], but accounting for such a complex and variable process is difficult. If all the strains under study are closely related, for example when studying a local outbreak, then most recombination among them will have little effect compared to imported sequences from other lineages, which would introduce a relatively high number of nucleotide changes. Such imports from other lineages can be inferred from the presence of multiple substitutions in a short genomic region and on a single phylogenetic branch [54,55]. However, recombination events for which both donor and recipient belong to the population under study can be sometimes detected if different genome regions support different phylogenetic topologies [49,56,57]. In addition to its potential to distort phylogenetic inference, recombination can also create a false signal of apparent mutational evolution by introducing additional divergence between samples taken at different time points [58]. There are thus several important reasons to test and account for recombination in the analysis of measurably evolving pathogens, and improved methods are needed for handling it.

Variation in genome content

DNA virus and bacterial genomes can exhibit significant variation in gene content and order, even between close relatives. For a given population, the ‘core genome’ represents the regions found in all the genomes, whereas the regions found in at least one but not necessarily all genomes are called the ‘pan-genome’ [59]. Since standard approaches to studying measurably evolving populations (Box 3) use alignments of homologous sequences, they can apply only to the core genome. The presence of paralogous genetic elements complicates the generation of such alignments. One approach is to find the genes homologous in all genomes, and align them separately [60]. Another approach is to align each genome against a reference,

typically via reference-based assembly of short read sequence data [61]. Reference-free alignment of all genomes against each other is also possible [62,63], but only for relatively small numbers of taxa. Although the core genome provides a convenient starting point for the analysis of measurably evolving pathogens, it ignores gene content variation that can be of critical importance [64]. For example, in an analysis of *Salmonella enterica* serovar Agona over several decades, most genetic diversity was due to the gain and loss of several bacteriophages, plasmids and integrative conjugate elements [65]. Gain and loss of genomic regions occur concurrently with diversification of the core genome via mutation and homologous recombination, but the same elements can be gained or lost multiple times, hence trees based on gene content can look very different from those based on the core genome [57,66]. There is currently no well-established framework for studying gene content variation of measurably evolving populations, making this another important area in need of methodological development. Revisiting earlier approaches, such as mathematical models for phylogenetic inference based on gene order [67], might be fruitful in this context.

Conclusions

As whole-genome sequencing promises to render most, or even all, microbial pathogens measurably evolving, there is a critical need for increased scientific dialogue among evolutionary biologists, epidemiological modellers and microbiologists, with a shared aim of developing methods that can accommodate the specific biological complexities inherent in many bacterial and virus systems. Box 4 summarises some of the outstanding research questions that, in our view, warrant particular attention. Only by overcoming these new and long-standing problems will

the exciting advances currently being made in microbial evolution reach their full potential.

Acknowledgements

This paper benefitted from the discussions during a workshop at the University of Glasgow in 2013, funded by the RAPIDD programme of the Science and Technology Directorate of the US Department of Homeland Security and National Institutes of Health Fogarty International Center. We would like to thank the workshop participants for the contributions made during the workshop. We also thank Allen Rodrigo and one anonymous reviewer for their constructive comments on an earlier version of this paper. R.B. is supported by NIH grant RO1 AI047498 and BBSRC grant BB/L010569/1. J.L.-S. is funded by NSF grants EF-0928690 and OCE-1335657. O.G.P. received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 614725-PATHPHYLODYN. X.D. would like to acknowledge the NIHR for Health Protection Research Unit funding.

References:

1 Drummond, A.J. *et al.* (2003) Measurably evolving populations. *Trends Ecol. Evol.* 18, 481–488

2 Volz, E.M. *et al.* (2009) Phylodynamics of infectious disease epidemics. *Genetics* 183, 1421–1430

3 Volz, E.M. *et al.* (2013) Viral Phylodynamics. *PLoS Comput. Biol.* 9, e1002947

4 Grenfell, B.T. *et al.* (2004) Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* 303, 327–332

5 Pybus, O.G. and Rambaut, A. (2009) Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10, 540–550

6 Kao, R.R. *et al.* (2014) Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends Microbiol.* 22, 282–291

7 Didelot, X. *et al.* (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* 13, 601–612

8 Yoshida, K. *et al.* (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife* 2, e00731

9 Firth, C. *et al.* (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.* 27, 2038–2051

10 Li, Y. *et al.* (2007) On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15787–15792

11 Kerr, P.J. *et al.* (2012) Evolutionary History and Attenuation of Myxoma Virus on Two Continents. *PLoS Pathog.* 8, e1002950

12 Bryant, J.M. *et al.* (2013) Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: A retrospective observational study. *Lancet Respir. Med.* 1, 786–792

13 Ford, C.B. *et al.* (2011) Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nat. Genet.* 43, 482–486

14 Biek, R. *et al.* (2012) Whole Genome Sequencing Reveals Local Transmission Patterns of Mycobacterium bovis in Sympatric Cattle and Badger Populations. *PLoS Pathog.* 8, e1003008

15 Stack, J.C. *et al.* (2013) Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. R. Soc. B Biol. Sci.* 280, 20122173

16 Worby, C.J. *et al.* (2014) Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. *PLoS Comput. Biol.* 10, e1003549

17 Nakazawa, Y. *et al.* (2013) Phylogenetic and ecologic perspectives of a monkeypox outbreak, Southern Sudan, 2005. *Emerg. Infect. Dis.* 19, 237–245

18 Poon, A.F.Y. *et al.* (2012) Reconstructing the Dynamics of HIV Evolution within Hosts from Serial Deep Sequence Data. *PLoS Comput. Biol.* 8, e1002753

19 Ypma, R.J.F. *et al.* (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195, 1055–1062

20 Morelli, M.J. *et al.* (2012) A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Comput. Biol.* 8, e1002768

21 Jombart, T. *et al.* (2014) Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Comput. Biol.* 10, e1003457

22 Ho, S.Y.W. and Larson, G. (2006) Molecular clocks: when times are a-changin'. *Trends Genet.* 22, 79–83

23 Comas, I. *et al.* (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat. Genet.* 45, 1176–1182

24 Ho, S.Y.W. *et al.* (2011) Time-dependent rates of molecular evolution. *Mol. Ecol.* 20, 3087–3101

25 Duchêne, S. *et al.* (2014) Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.* 281, 20140732

26 Schuenemann, V.J. *et al.* (2013) Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science* 341, 179–183

27 Wagner, D.M. *et al.* (2014) Yersinia pestis and the Plague of Justinian 541-543 AD: A genomic analysis. *Lancet Infect. Dis.* 14, 319–326

28 Bos, K.I. *et al.* (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* 514, 494–497

29 Smith, T.F. *et al.* (1988) The phylogenetic history of immunodeficiency viruses. *Nature* 333, 573–575

30 Holmes, E.C. (2003) Molecular clocks and the puzzle of RNA virus origins. *J. Virol.* 77, 3893–3897

31 Wertheim, J.O. and Kosakovsky Pond, S.L. (2011) Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* 28, 3355–3365

32 Wertheim, J.O. *et al.* (2013) A case for the ancient origin of coronaviruses. *J. Virol.* 87, 7039–45

33 Wertheim, J.O. *et al.* (2014) Evolutionary Origins of Human Herpes Simplex Viruses 1 and 2. *Mol. Biol. Evol.* 31, 2356–2364

34 Duffy, S. *et al.* (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276

35 Chattopadhyay, S. *et al.* (2009) High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12412–12417

36 Lee, H. *et al.* (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2774–E2783

37 Uphoff, S. *et al.* (2013) Single-Molecule DNA Repair in Live Bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 110, 8063–8068

38 Martincorena, I. *et al.* (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485, 95–98

39 Cui, Y. *et al.* (2013) Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc. Natl. Acad. Sci. U. S. A.* 110, 577–582

40 Jolivet-Gougeon, A. *et al.* (2011) Bacterial hypermutation: Clinical implications. *J. Med. Microbiol.* 60, 563–573

41 Lieberman, T.D. *et al.* (2014) Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* 46, 82–87

42 Colangeli, R. *et al.* (2014) Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLoS One* 9, e91024

43 Didelot, X. *et al.* (2012) Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biol.* 13, R118

44 Eyre, D.W. *et al.* (2013) Diverse Sources of C. difficile Infection Identified on Whole-Genome Sequencing. *N. Engl. J. Med.* 369, 1195–1205

45 Didelot, X. and Maiden, M.C.J. (2010) Impact of recombination on bacterial evolution. *Trends Microbiol.* 18, 315–322

46 Everitt, R.G. *et al.* (2014) Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat. Commun.* 5, 3956

47 Croucher, N.J. *et al.* (2013) Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* 45, 656–663

48 Yahara, K. *et al.* (2014) Efficient inference of recombination hot regions in bacterial genomes. *Mol. Biol. Evol.* 31, 1593–1605

49 Ansari, M.A. and Didelot, X. (2014) Inference of the Properties of the Recombination Process from Whole Bacterial Genomes. *Genetics* 196, 253–265

50 Cordero, O.X. *et al.* (2012) Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science* 337, 1228–1231

51 Sheppard, S.K. *et al.* (2013) Progressive genome-wide introgression in agricultural Campylobacter coli. *Mol. Ecol.* 22, 1051–1064

52 Qin, L. and Evans, D.H. (2014) Genome scale patterns of recombination between coinfecting vaccinia viruses. *J. Virol.* 88, 5277–5286

53 Hedge, J. and Wilson, J. (2014) Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* 5, e02158–14

54 Didelot, X. and Falush, D. (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175, 1251–1266

55 Croucher, N.J. *et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331, 430–434

56 Didelot, X. *et al.* (2010) Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* 186, 1435–1449

57 Shapiro, B.J. *et al.* (2012) Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* 336, 48–51

58 Sánchez-Busó, L. *et al.* (2014) Recombination drives genome evolution in outbreak-related Legionella pneumophila isolates. *Nat. Genet.* 46, 1205–1211

59 Vernikos, G. *et al.* (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23, 148–154

60 Maiden, M.C.J. *et al.* (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* 11, 728–736

61 Nielsen, R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451

62 Darling, A.E. *et al.* (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS One* 5, e11147

63 Angiuoli, S. V. and Salzberg, S.L. (2011) Mugsy: Fast multiple alignment of closely related whole genomes. *Bioinformatics* 27, 334–342

64 Croll, D. and McDonald, B. a (2012) The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* 8, e1002608

65 Zhou, Z. *et al.* (2013) Neutral Genomic Microevolution of a Recently Emerged Pathogen, Salmonella enterica Serovar Agona. *PLoS Genet.* 9, e1003471

66 Didelot, X. *et al.* (2009) Inferring genomic flux in bacteria. *Genome Res.* 19, 306–317

67 Bourque, G. and Pevzner, P.A. (2002) Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.* 12, 26–36

68 Rodrigo, A.G. and Felsenstein, J. (1999) Coalescent approaches to HIV population genetics. *The evolution of HIV, 233-272.*

69 Drummond, A.J. *et al.* (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* 22, 1185–1192

70 Orlando, L. and Cooper, A. (2014) Using Ancient DNA to Understand Evolutionary and Ecological Processes. *Annu. Rev. Ecol. Evol. Syst.* 45, 573–598

71 Ypma, R.J.F. *et al.* (2012) Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B Biol. Sci.* 279, 444–450

72 Mollentze, N. *et al.* (2014) A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc. R. Soc. B Biol. Sci.* 281, 20133251

73 Harris, S.R. *et al.* (2013) Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. *Lancet Infect. Dis.* 13, 130–136

74 Didelot, X. *et al.* (2014) Bayesian inference of infectious disease transmission from whole genome sequence data. *Mol. Biol. Evol.* 31, 1869–1879

75 Mutreja, A. *et al.* (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477, 462–465

76 Harris, S.R.R. *et al.* (2010) Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* 327, 469–474

77 Walker, T.M. *et al.* (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect. Dis.* 13, 137–146

78 Mathers, A.J. *et al.* (2015) Klebsiella pneumoniae carbapenemase (KPC) producing K. pneumoniae at a Single Institution: Insights into Endemicity from Whole Genome Sequencing. *Antimicrob. Agents Chemother.*, AAC.04292-14

79 Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214

80 Drummond, A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4, e88

81 Chewapreecha, C. *et al.* (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* 46, 305–309

82 He, M. *et al.* (2013) Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nat. Genet.* 45, 109–113

83 Ho, S.Y.W. (2014) The changing face of the molecular evolutionary clock. *Trends Ecol. Evol.* 29, 496–503

84 Ho, S.Y.W. and Duchêne, S. (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* 23, 5947–5965

85 Zhou, Z. *et al.* (2014) Transient Darwinian selection in Salmonella enterica serovar Paratyphi A during 450 years of global spread of enteric fever. *Proc. Natl. Acad. Sci. U. S. A.* 111, 12199–12204

Box 1 - A primer on measurably evolving populations

When a population is genetically sampled, researchers commonly assume that the accumulation of evolutionary events over the time scale of sampling is negligible. As Drummond et al. pointed out [1], this assumption might not always hold. Consider two lineages that share a recent common ancestor and that are sampled at different points in time (such data are termed "serially-sampled" or "time-stamped"). When comparing the respective genetic divergence of each from their most recent common ancestor (MRCA), the lineage sampled earlier is expected to be less divergent since it had less time to accumulate substitutions compared to the lineage sampled later (Figure Ia). This additional divergence (δ) is a product of the evolutionary rate per site per unit time (μ), the duration of the sampling interval (t) and the number of sites in the sequence considered (l), so $\delta = \mu t l$. When δ is not distinguishable from zero, sampled sequences can be considered isochronous (i.e. evolutionarily equivalent to sequences that were sampled at precisely the same time). In contrast, δ>0 implies a measurably evolving population and samples described as heterochronous, indicating that the differences in their sampling date must be taken into account during evolutionary analysis. This is often accomplished by using phylogenetic molecular clock models that constrain the distance of each tip from the tree root to be proportional to its sampling date (Figure Ib).

By rescaling genealogies into natural units of time, the concept of measurably evolving populations is pertinent to a suite of research areas at the interface of population biology and molecular evolution, including phylogenetics, demographic modelling, palaeobiology, epidemiology and phylogeography. Combined with coalescent-based inference methods [68], it has contributed to our understanding of how historical populations changed through time [69]. It has been frequently applied to ancient DNA from eukaryotic species (for which the sampling interval t is large) [70], and to RNA viruses (for which the mutation rate μ is large) [69]. Specifically, the

concept has provided insights into the emergence and transmission of major pathogens including HIV, influenza and Ebola virus [3,5]. Similar advances might now be possible for a much wider range of organisms through the generation of whole genome data, thus capitalising on the third factor that can make populations measurably evolving, the number of sites in the sampled sequence (l).

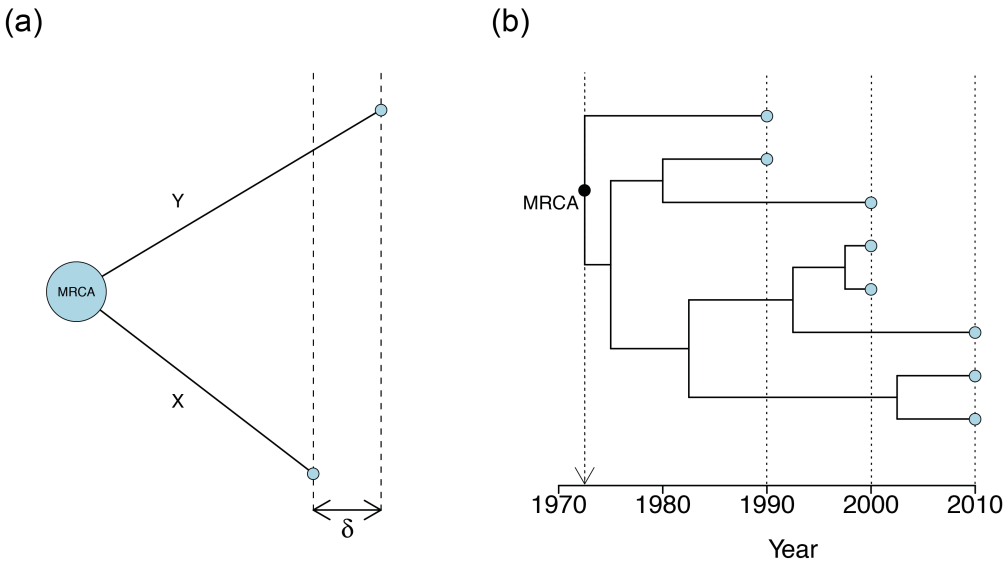


Figure I: a) Populations are considered measurably evolving if two lineages (X and Y) sampled at separate time points, are different with respect to their genetic divergence from their most recent common ancestor (MRCA), and that this difference is statistically greater than zero. Figure redrawn from [1]. b) A genealogy with dated tips in which branch lengths have been estimated using a phylogenetic molecular clock model. Such genealogies provide estimated dates for all internal nodes, including the MRCA (arrow).

Box 2: Integrating genetic and epidemiological data

The minimum data required for studying measurably evolving pathogens consists of sampled gene sequences and their relevant sampling dates. For many DNA viral and bacterial pathogens, the time scale over which novel genomic variation is observed is of the same order or even slower than the time scale of transmission from host-to-host (Figure 2). In this case, resolving 'who infected whom' cannot be done solely using molecular information, which has led to several recent modelling efforts to integrate genetic with epidemiological data (Figure I). The additional epidemiological information can be geographical, for example, when considering infections that spread from one place to another as in the case of foot-and-mouth disease virus spreading between farms. The greater propensity for transmission between farms in spatial proximity can be accounted for by an additional term in the likelihood function that penalizes distant transmission events [20,71]. Analysing such space-time-genetic data is also possible in endemic regions where only a fraction of incident cases are observed and multiple introductions of the pathogen might have occurred [72]. Additional epidemiological information is sometimes available at the individual host level, further constraining the set of transmission trees that are consistent with the data. For example an estimate of when a host became infectious implies that transmission from that host could not have happened beforehand [19,20]. In the study of nosocomial infections, detailed information is typically available about patients' admission and release dates, whereabouts in hospitals, symptoms and treatments. Current research aims to analyze this information jointly with pathogen genetic data [44,73].

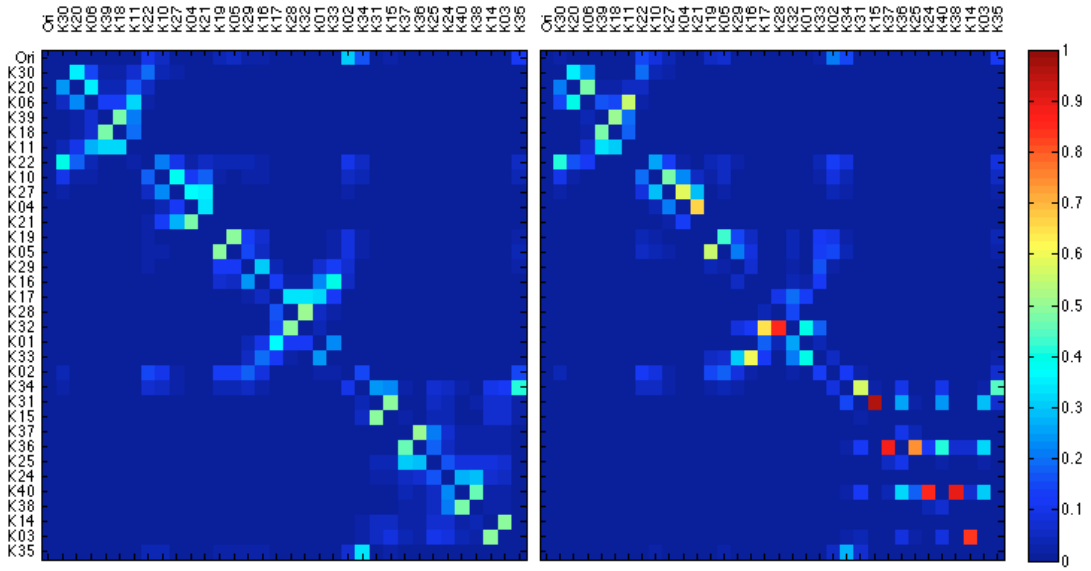


Figure I. Example of inferring transmission links using a Bayesian approach that integrates both genetic and epidemiological data. The data are taken from a recent study of a tuberculosis outbreak in Canada [74]. The two matrices show the posterior probability of transmission from one individual (row) to another (column), with warmer colours representing higher probabilities. The matrix on the left is based on the genomic data only, whereas the matrix on the right is based on both genomic and epidemiological data, namely geographic data and indications of infectiousness provided by smear and skin tests. Accounting for these additional data significantly reduces the uncertainty in who infected whom.

Box 3: Methods for handling time-stamped genomic data

A common first step in the exploration of heterochronous sequence data is to construct a rooted phylogeny and perform a linear regression between the sampling dates of each sequence and their corresponding root-to-tip genetic distances (Figure Ia). A strong positive linear relationship between time and genetic divergence indicates that the sequences contain 'temporal signal' and are suitable for analysis using molecular clock models. Further, it provides preliminary information about the rate of molecular evolution and the date of the most recent common ancestor of the sample (Figure Ia). This method has been used successfully for many bacterial pathogens, for example *Vibrio cholerae* [75], *Streptococcus pneumoniae* [55] and *Staphylococcus aureus* [76]. A related technique can be used when two pathogen sequences have been sampled longitudinally from the same host. When many such pairs are available, a linear regression can be applied to the genetic distance between each sequence pair and the time between first and second sampling (Figure Ib). This reveals both the rate of molecular evolution and the level of intra-host genetic diversity. This method has been applied to *Clostridium difficile* [43], *Mycobacterium tuberculosis* [77] and *Klebsiella pneumoniae* [78]. While providing a useful starting point, these regression approaches are limited as estimation tools because the data points are non-independent (due the presence of shared common ancestry or multiple pairwise comparisons) and because they are based on a point estimate of phylogeny. These limitations can be overcome with Bayesian phylogenetic approaches (e.g. those implemented in BEAST [79]) that can co-estimate evolutionary rates and serially-sampled genealogies (see Box 1). Such analysis requires that a molecular clock model is chosen, the simplest of which is the 'strict clock' that assumes that all phylogeny branches evolve at the same rate. However, there are many reasons to suspect that evolutionary rates within microbial

species vary through time or among lineages (see main text). In such cases it might be more appropriate to use a 'relaxed' molecular clock model [80], that allows among-branch variation in evolutionary rates, for example the uncorrelated relaxed clock model that has been applied successfully in several bacterial genomic studies [39,47,81,82]. For an overview of current molecular clock models see [83,84]. One limitation of phylogenetic methods such as BEAST is that they do not account for recombination, which can disrupt the temporal signal (see main text). A pragmatic solution is to use a software that can detect recombination such as ClonalFrame [54] and apply BEAST only to the data that were not affected by recombination [58,85].

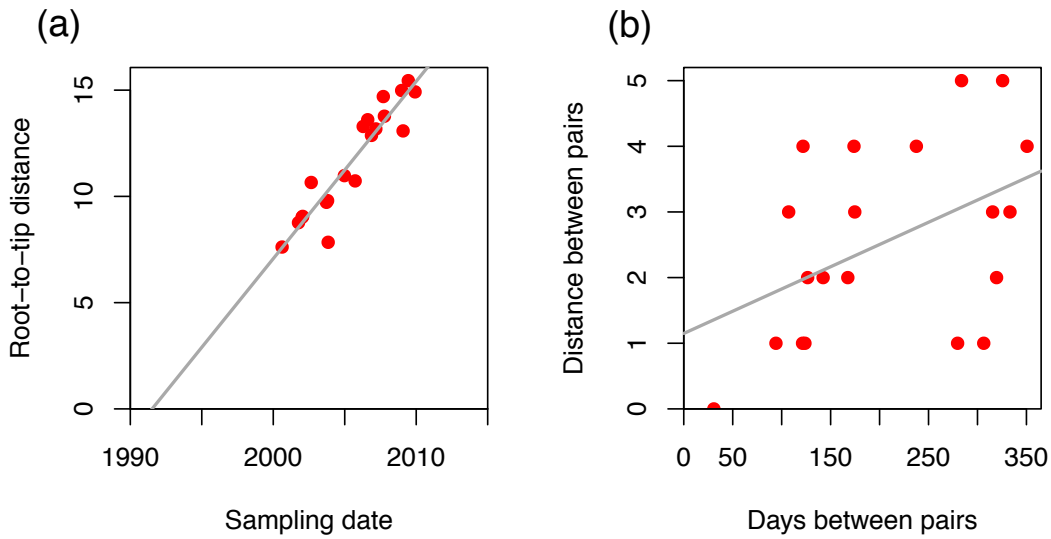


Figure I. Two types of linear regression analysis based on heterochronous genetic sequences. (a) From a rooted phylogeny, root-to-tip distances are shown on the y-axis and sampling dates on the x-axis. The slope is an estimate of the rate of molecular evolution, and the x-intercept corresponds to the estimated date of the root. (b) From pairs of genomes sampled sequentially from the same hosts, the distances between sequences are shown on the y-axis and the sampling intervals on the x- axis. The slope is an estimate of the rate of molecular evolution, and the y-intercept corresponds to the average distance between pairs of sequences sampled at the same time.

Box 4: Outstanding Questions

Pathogen genomes often contain genetic polymorphisms other than SNPs, such as indels and plasmids. Do these polymorphisms also fit the 'measurable evolution' paradigm and can they be informative about transmission processes?

How can analyses incorporate genetic variation outside the core genome and consider variation in genome content?

How can genomic and non-genomic data be integrated most effectively to quantify transmission links across different scales?

Does the concept of measurably evolving populations also apply to eukaryotic pathogens (e.g. fungi, protozoa) on time scales that are relevant to transmission?

What are the mechanisms causing time-dependent rates of molecular evolution and how universal are they? What are the most appropriate analytical frameworks for dealing with this time-dependency?

What are the underlying processes causing evolutionary rate heterogeneity within pathogen species and how are they best accounted for during sequence analyses? Could processes and their relative contributions be inferred from genetic patterns?

How is recombination best detected and accounted for in the evolutionary analysis of microbial genomes?

Glossary

DNA virus: viruses that encode their genetic material as DNA. Double-stranded (ds)DNA viruses use host enzymes to replicate their genomes. Due the proof-reading activity in these replicases, mutational change per replication event tends to be rare. In contrast, single-stranded (ss)DNA viruses can evolve at higher rates similar to those seen for RNA viruses.

Evolutionary rate: estimated rate at which nucleotide changes (per site or per genome) are observed within a population sampled over time. Sometimes used synonymously with substitution rate, which technically refers to the rate at which these nucleotide changes become fixed at the population level (also see Mutation rate). Unless indicated otherwise we generally refer in this article to rates per unit time rather than per generation since information on generation time for natural transmission is often unavailable.

Heterochronous: refers to sequence data sampled over sufficiently long time periods to allow measurable evolution between sampling times. In contrast, sequences for which no such effect is detectable and which can be considered to have been sampled at effectively the same time point, are termed isochronous (see Box 1).

Mutation rate: rate at which novel mutations arise (per site or per genome), most of which are subsequently removed by purifying selection. This rate therefore represents the upper biological limit for the amount of genetic change per unit of time (see Evolutionary rate).

Phylodynamics: scientific discipline that aims to infer the population processes that gave rise to particular phylogenetic patterns, as identified from genetic sequence data. Often, but not exclusively, applied in the context of infectious disease transmission, phylodynamic approaches have been used to study processes including immune selection, population expansion, spatial movement, and transmission or recovery rates.

RNA virus: group of viruses that use RNA as their genetic material and produce their own enzyme (an RNA polymerase) for genome replication. Due short generation times (often days) and a lack of proof-reading capability in the polymerase, many RNA viruses quickly accumulate genetic changes at the genome level (also see: DNA virus).

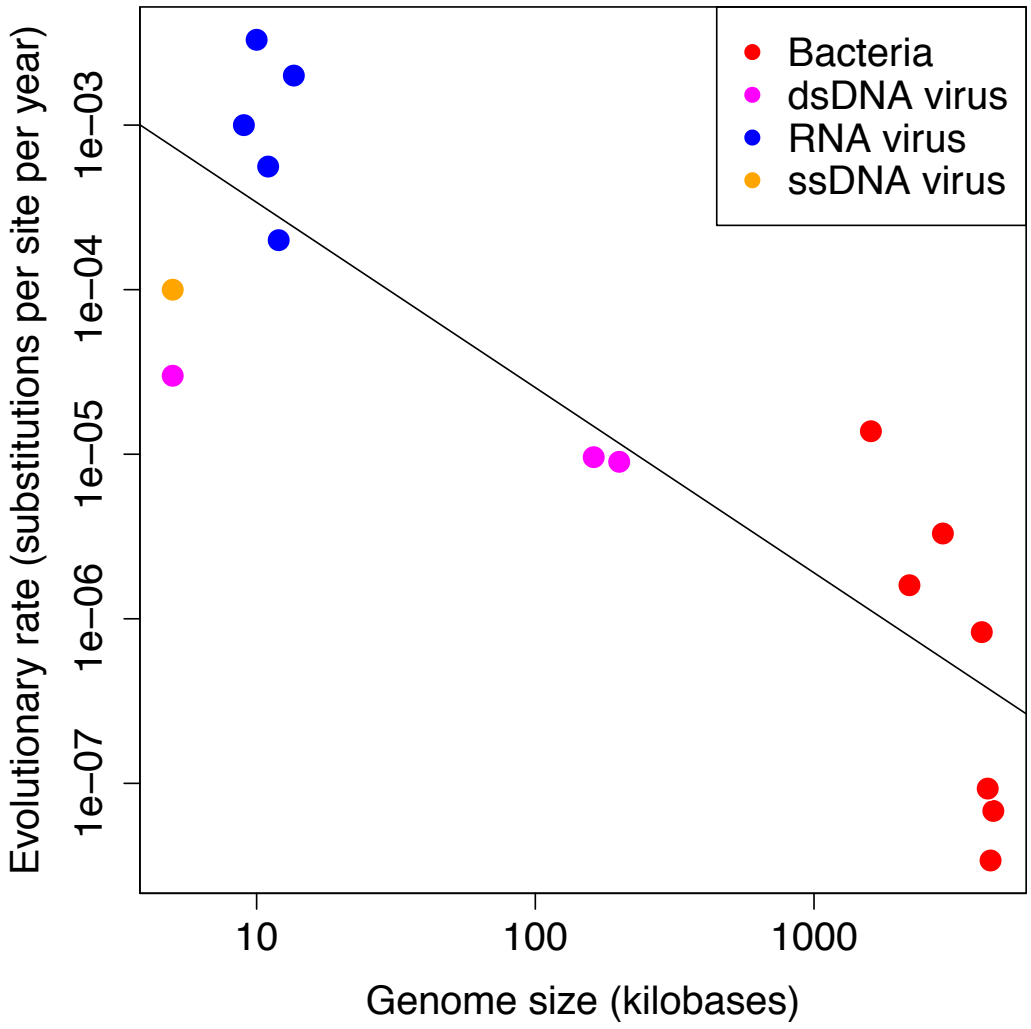
Figure legends

Figure 1. There is a broad negative relationship between evolutionary rate and genome size across a range of different viruses and bacteria. Evolutionary rates shown are based on a representative selection of published datasets of heterochronously sampled, complete or partial genomes sampled between one and six decades apart. See Supplementary Table S1 for details on rate estimates.

Figure 2. The relative time scales of epidemiological and evolutionary processes (at the whole genome level) can vary widely among viral and bacterial pathogens. Average intervals between transmission and nucleotide substitution events were calculated as the reciprocal of the pathogen's reported generation time and estimated evolutionary rate, respectively. Evolutionary rates were estimated based on published datasets of heterochronous genomes sampled up to two decades apart. Axes are on a log scale but due to considerable uncertainties and heterogeneities associated with the underlying parameters, are only labeled with broad temporal units. For pathogens above the unity line, novel genetic variation is expected to become fixed faster than the average time between host-to-host transmission events, making it possible in principle to reconstruct individual transmission pathways from genomic data; the same is not true for pathogens below the unity line. The lower end of the evolutionary time scale is ultimately bounded by the underlying mutation rate per genome replication event, as indicated by the blunt left-hand sides of the clouds representing parameter estimates. See Supplementary Table S1 for details on rate estimates.

Figure 3. Consistent with a general pattern for measurably evolving populations, the evolutionary rates of microbial pathogens decrease as a function of the time span over which they are estimated. Data shown are selected representative examples, including one group of RNA viruses (primate lentiviruses SIV and HIV in blue, taken from [25], genomic rates extrapolated from *pol* gene sequences) and several bacterial pathogens. See Supplementary Table S1 for details.

Figure 1


Bacteria
dsDNA virus
RNA virus
ssDNA virus
Evolutionary rate (substitutions per site per year)
1e−03
1e−04
1e−05
1e−06
1e−07
10
100
1000
Genome size (kilobases)

5

Figure 2

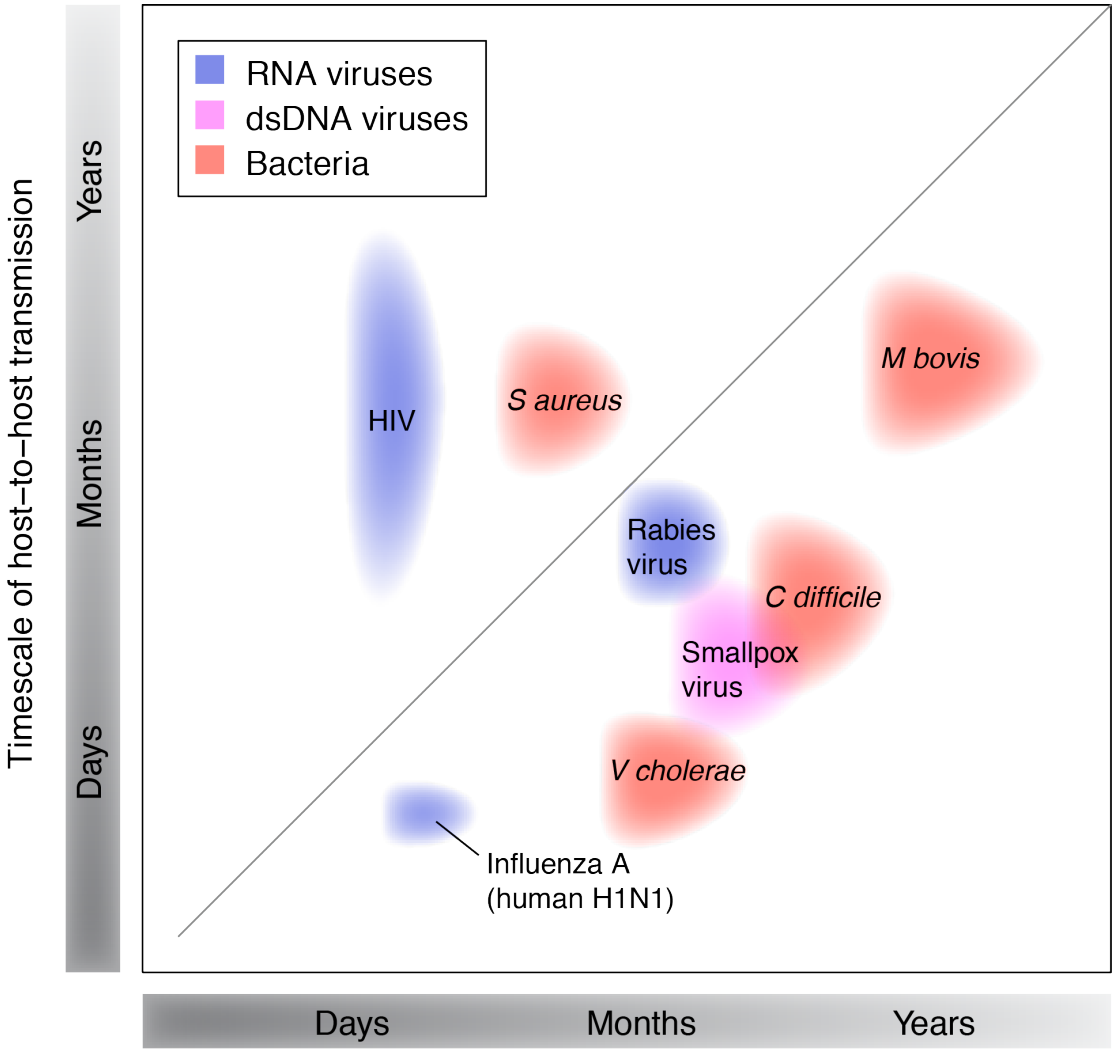
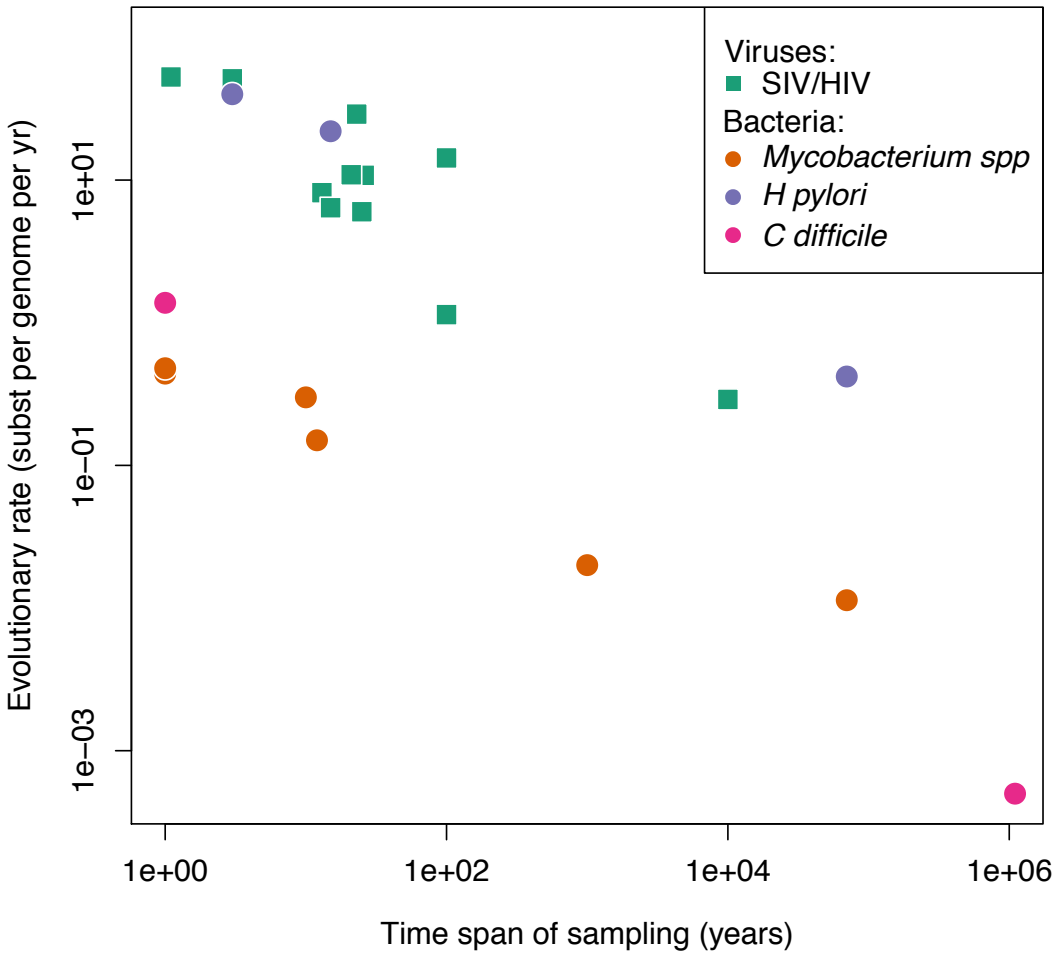

RNA viruses
dsDNA viruses
Bacteria
HIV
S aureus
M bovis
Rabies virus
C difficile
Smallpox virus
V cholerae
Influenza A (human H1N1)
Timescale of host–to–host transmission
Years
Months
Days
Days
Months
Years
Timescale at which novel genomic variation is observed

Figure 3


Viruses:
SIV/HIV
Bacteria:
Mycobacterium spp
H pylori
C difficile
Evolutionary rate (subst per genome per yr)
1e+01
1e−01
1e−03
1e+00
1e+02
1e+04
1e+06
Time span of sampling (years)